Mortality prediction in the intensive care: Role of mathematical models in benchmarking and decision-making
Minne, L.

Citation for published version (APA):
Minne, L. (2013). Mortality prediction in the intensive care: Role of mathematical models in benchmarking and decision-making
Chapter 6

Statistical process control for monitoring standardized mortality ratios of a classification tree model

Minne L, Eslami S, de Keizer NF, de Jonge E, de Rooij SE, Abu-Hanna A.
Statistical process control for monitoring standardized mortality ratios of a classification tree model.
The original publication is available at http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/methods/contents/archive/issue/1566/manuscript/17844.html
Abstract

Objectives: The ratio of observed to expected mortality (standardized mortality ratio, SMR), is a key indicator of quality of care. We use PreControl Charts to investigate SMR behaviour over time of an existing tree-model for predicting mortality in intensive care units (ICUs) and its implications for hospital ranking. We compare the results to those of a logistic regression model.

Methods: We calculated SMRs of 30 equally-sized consecutive subsets from a total of 12,143 ICU patients aged 80 years or older and plotted them on a PreControl Chart. We calculated individual hospital SMRs in 2009, with and without repeated recalibration of the models on earlier data.

Results: The overall SMR of the tree-model was stable over time, in contrast to logistic regression. Both models were stable after repeated recalibration. The overall SMR of the tree on the whole validation set was statistically significantly different (SMR 1.00 ± 0.012 vs. 0.94 ± 0.01) and worse in performance than the logistic regression model (AUC 0.76 ± 0.005 vs. 0.79 ± 0.004; Brier score 0.17 ± 0.012 vs. 0.16 ± 0.010). The individual SMRs’ range in 2009 was 0.53–1.31 for the tree and 0.64–1.27 for logistic regression. The proportion of individual hospitals with SMR >1, hinting at poor quality of care, reduced from 38% to 29% after recalibration for the tree, and increased from 15% to 35% for logistic regression.

Conclusions: Although the tree-model has seemingly a longer shelf life than the logistic regression model, its SMR may be less useful for quality of care assessment as it insufficiently responds to changes in the population over time.

Introduction

The standardized mortality ratio (SMR) is a key indicator in quality of care assessment programs [1, 2] enabling risk-adjusted comparison of hospitals [3]. The SMR is the ratio of the observed number of deaths to the (case mix-adjusted) predicted number of deaths [4]. There are two uses of the SMR. The first pertains to calculating the global SMR of the whole group of patients from all units, while the second pertains to calculating the individual SMR of each unit.

The global SMR provides a measure of calibration of the underlying prognostic model (i.e. the model used to obtain the predicted number of deaths). For a well-calibrated model the global SMR is expected to be 1 regardless of the dataset for which prediction is performed. To illustrate, models developed using logistic regression will always result in SMR = 1 on the developmental dataset itself, but not necessarily on other validation sets. This global SMR will tend to be closer to 1 in the immediately prospectively collected validation sets than those collected later in time. When time elapses, the SMR can be influenced by changes over time in both the quality of care delivered and in patient mix. Values deviating from 1 indicate poorer global calibration. If a model corrects perfectly for patient mix, however, a change in the population without a change in the quality of care should still result in a global SMR of 1. In this case, a global SMR ≠ 1 would indicate a change in quality of care.

The individual SMR is used to regularly rank the units among themselves against the reference standard (i.e. the global SMR) [5]. In this setting SMR = 1 indicates that the delivered quality of care of that unit is equivalent to the expected quality of care of the “standard unit”; SMR >1 indicates poorer than expected quality of care, and SMR <1
indicates better than expected quality of care. If individual SMRs are calculated on the basis of a poorly calibrated model (global SMR ≠ 1) however, they will provide an incorrect view of the delivered quality of care.

Although there is much work on developing and internally validating prognostic models [6] there is very little work on monitoring SMR behavior and its implications over time, especially for non-parametric models. There are several reasons that limit the “shelf-life” of prognostic models, including population drift and the introduction of new medical technologies [7]. A stable SMR over time can be indicative of stability of the population and environment, or, worse, of the non-responsiveness of the model to changes over time. Models are useful if they are sufficiently responsive to changes over time, but they should be timely recalibrated to reflect these changes.

In this paper we use a statistical process control (SPC) method to investigate the SMR behavior over time and its implications on ranking hospitals for an already published tree-model, and compare the results to those obtained by a logistic regression predictive model. SPC originates from industry, but its graphical and statistical orientation for studying variations over time formed our motivation for using it in monitoring SMR behavior and analyzing its implication on hospital ranking.

Methods
In this prospective cohort study, we monitored the SMR of a tree-model over the course of time by partitioning the prospective data in 30 time-ordered equally-sized and mutually exclusive groups and calculating the global SMR for each group. These SMRs were then scrutinized by the SPC-method described below, and then compared to the results based on the logistic regression model. The results of the logistic regression model were obtained in the same way as for the tree-model and are described in [8]. All statistical analyses were conducted in the R statistical environment [9]. We tested for statistical significant differences between the models by calculating differences in their performance in 1000 bootstrap samples and checking whether the interval between the 2.5 and 97.5 percentiles of the 1000 differences does not include zero. Exact p-values were obtained from the bootstrap distribution of the differences centered around their mean.

Model Development and Internal Validation
In 2007 we developed a classification tree-model to predict mortality risks for Intensive Care Unit (ICU) patients aged 80 years and older, and compared its predictive performance to that of a logistic regression model [10]. The dataset consisted of 6,867 ICU admissions of 21 ICUs between January 1997 to December 2003, of which two thirds (N = 4,578) was randomly selected for development and one third (N = 2,289) for internal validation of both models. The patient characteristics are described in [10].

The tree-model was built by means of binary recursive partitioning (RPA) using the CART algorithm [11]. The tree structure and underlying data variables are shown in Figure 6.1. The splits in the tree are based on information gain. And the tree was completely over-grown to overfit the data. Then the tree was pruned back based on the complexity that resulted in the minimum 10 cross-validated error. The cross-validation error will usually first decrease with tree size, then reach a minimum which is associated with the optimal tree size, but then start increasing again due to overfitting. The package used for RPA was Rpart, which is an implementation of CART in the R statistical environment [9].
The performance of the tree-model in the internal validation set in terms of the area under the receiver operating characteristic curve (AUC) [12] and Brier score (± standard deviation) [13] were 0.77 ± 0.01 and 0.16 ± 0.005, respectively.

Figure 6.1. Classification tree-model for predicting mortality in ICU patients aged 80 years and older as reported in [6].

The logistic regression model was obtained by first-level recalibration of the Simplified Acute Physiology Score-II (SAPS-II) model on the developmental set. The SAPS-II score is the sole covariate in the model and it is based on a set of 12 physiological variables (e.g., age, type of admission: scheduled surgical, unscheduled surgical, or medical) and underlying disease variables (i.e., AIDS, metastatic cancer and hematological malignancy) collected in the first 24 hours of admission [14]. The SAPS-II score is a commonly used measure of severity of illness in the ICU. First-level recalibration refers to fitting a new logistic regression model using the outcome from the new dataset and the (log odds of the) original probabilities as the sole input variable. The recalibrated SAPS-II model (i.e., rSAPS-II) had the following linear predictor (LP): \(-3.623 + 0.073 \times \text{SAPS-II} - 0.089 \times \log(\text{SAPS-II} + 1)\). The probability of death is simply \(\exp(\text{LP})/\exp(1 + \text{LP})\). The performance of rSAPS-II in the internal validation set in terms of the AUC and Brier score (± standard deviation) were 0.77 ± 0.01 and 0.16 ± 0.01, respectively.

**Statistical Process Control**

SPC is a rigorous time series analysis and graphical data presentation (often yielding insights into the data more quickly and in a more understandable way than other statistical techniques) that can identify structural changes in a process [15]. Its primary tool, the process control chart, is a plot of the data over time with three additional lines; the center line (usually the mean) and upper and lower control limits, typically set at ±3 adjusted...
standard deviations (called sigma) from the mean. When the process has only inherent
variation then the data points exhibit no special patterns and are within the control limits.
The process is then said to be “in control” or “stable” as the process is predictable within
certain limits. Special cause variation, on the other hand, refers to variation caused by a
structural, external change in the process, such as a change in quality of care. This signifies
that the process is no longer stable or predictable and has changed, either for better
or worse [16].

One type of control chart is the PreControl Chart. While the limits of control charts
are usually data driven, PreControl Charts distinguish themselves from other charts by
allowing users to pre-specify the limits of multiple zones [17]. As described in [8], we
defined these zones by mean $\pm 2$ standard deviations (green/safe zone), 4 standard devi-
ations (yellow/warning zone) and 6 standard deviations (red/critical zone) of the moni-
tored statistic, here the SMR. For each of the 30 consecutive equally-sized subsets of size
405, we computed the global SMR and its 95% confidence interval based on 1000 boot-
strap samples. Confidence intervals were only shown if the SMR was significantly above
or below 1.0. The number of groups was chosen to be 30 because it still results in sizable
groups and long enough series of performance measures to be scrutinized over time. Any
number between 12 and 36 is acceptable in SPC. As is customary, we considered the pro-
cess unstable when either two consecutive values fell in the yellow zone or one value fell
in the red zone. To calculate the limits we obtained the bootstrap sampling distribution of
the SMR in the internal validation set, based on 1000 bootstrap samples of size 405.

**Effects of Repeated Recalibration**

Similar to [8] we calculated the individual SMR of each ICU for the last year, 2009, based
on five recalibrated models, where a model’s predicted probabilities are updated based on
new data. Specifically, patients in 2009 were assigned to the original tree leaves, but the
mortality probability in a leaf was calculated as the mean mortality of the patients in that
1997–2007 and 1997–2008. Note that this is equivalent to first level recalibration as the
structure of the tree is not changed. To explore the effect of repeated recalibration on the
global SMR, we updated in the same way the probabilities for each time point $p$ in the 30-
point time series on the dataset from 1997 until the period just preceding $p$. Then the
model was prospectively evaluated on the dataset at time-point $p$.

**Case Study**

We used prospective data of all 12,143 consecutive admissions of 21 Dutch ICUs of pa-
tients aged 80 years and older between January 2004 and July 2009, which were the same
ICUs that participated in the developmental study of the tree-model. This is the same
dataset used for the previously published validation of the rSAPS-II model. Patient char-
acteristics are described in [8].

**Additional Analyses**

We conducted sensitivity analyses in which we changed the number (and hence size) of
the groups and stratified the analyses for medical and surgical admissions. In addition, we
compared our models to the logistic regression Acute Physiology And Chronic Health
Evaluation-II (APACHE-II) model and an alternative classification tree development
strategy. The APACHE-II model was recalibrated on the internal validation set in the same way as the rSAPS-II model. The alternative tree-model was obtained by pruning the tree not based on its minimal 10 cross-validated error, but based on the minimal error plus its standard deviation. This results in trees that are smaller in size than the original tree. Data of these additional analyses are not shown but addressed below.

Results
Overall performance of the tree-model on the prospective dataset in terms of the AUC (0.76 ± 0.005) and Brier score (0.17 ± 0.012) was statistically significantly worse (p = 0.0001) than the overall performance of rSAPS-II in this dataset (AUC = 0.79 ± 0.004, Brier score = 0.16 ± 0.010). The overall standard mortality ratio (SMR) of the tree-model (1.00 ± 0.012) was statistically significantly different (p = 0.0001) than that of rSAPS-II (0.94 ± 0.01). Figure 6.2 shows the PreControl Chart of the global SMR based on the tree-model.

![PreControl Chart of the global SMR based on the tree-model](image)

**Figure 6.2.** SMR of the original tree-model over time in subsets of the temporal validation set (N = 12,143). Each of the 30 consecutive time sets consists of 404 or 405 patient records. Means and standard deviations (sd) are based on the bootstrap sampling distribution based on 1000 bootstrap samples of the internal validation set (N = 2,289).

Only two non-consecutive points fall in the yellow zone, indicating that the SMR is stable. As shown by its 95% confidence interval, the SMR at point 11 was statistically significantly larger than 1. After repeated recalibration of the tree-model, only one point remains in the yellow zone of the PreControl Chart. In addition, none of the points are significantly higher or lower than 1. Contrary to these findings, the SMR based on the rSAPS-II model was instable, although this instability was adequately alleviated by repeated recalibration of the model [8]. The SMR of the rSAPS-II model was lower than the SMR based on the tree-model in 28 of the 30 time-points (starting at SMR rSAPS-II = 1.01 vs. SMR tree = 1.05 in the first time-point, and reducing to SMR rSAPS-II = 0.86 vs. SMR tree = 0.96 in the last one), and equal in time-points 4 (SMR = 1.00) and 7 (SMR = 1.09). The differences were statistically significant (p < 0.05) in 12 time-points. The recalibrated
APACHE-II model showed the same pattern as the rSAPS-II model (i.e. a run of the last 8 points below the mean; data not shown). Although the alternative tree-model generally had a higher error than the original one, it still remained stable over time (data not shown).

The implications of repeated recalibration can be adequately investigated by assessing the effects of recalibration on the SMRs of individual hospitals. The SMR of the tree-model without any recalibration ranged between 0.64 and 1.27, yielding 8 out of 21 hospitals (38%) in 2009 with SMR >1 (assessed as performing worse than average), and 13 (62% of hospitals) with SMR < 1 (assessed to deliver better care than average). Compared to the tree-model, the SMR of the rSAPS-II model, ranging from 0.53 to 1.31, was smaller in 16 out of 21 hospitals (76%). In four cases, the SMR was >1 for the tree-model while < 1 for rSAPS-II. After recalibration of the tree-model on data from 1997–2004, 1997–2005, 1997–2006, 1997–2007 and 1997–2008, the percentage of hospitals with SMR >1 reduced from 38% to 29% and the percentage with SMR < 1 increased from 62% to 71%. This is in contrast to the increase (from 15% to 35%) of hospitals with SMR >1 found for the recalibrated rSAPS-II model [8].

Discussion
The standardized mortality ratio (SMR) based on the tree-model showed a stable course over time, in contrast to the instability of the SMR found in the logistic regression model. Repeated recalibration positively affected the stability of the SMR of both models allowing them to better reflect changes in the patient population and treatment over time.

The SMRs of the tree-model were, on average, higher than the SMRs of the logistic regression model. This means that the quality of care is considered lower according to the tree-model than when employing the logistic regression model. After recalibration, the percentage of hospitals with SMR > 1 in 2009 reduced from 38% to 29% for the tree-model, and increased from 15% to 35% for the logistic regression model.

Although there are studies describing changes of SMR over time when comparing hospitals as a whole [18, 19], to our knowledge there are no other published studies that explore the implications on quality of care assessment of SMR behavior over the course of time, let alone for nonparametric models like classification trees. The SPC based strategy enables visualization and intuitive monitoring over time. The number of patients in the temporal validation set was large, and prospective data collection covered a period of six successive years. We conducted sensitivity analysis in which we changed the number (and hence size) of the groups and stratified the analyses for medical and surgical admissions. All yielded the same patterns (data not shown). In addition, we compared our models to another logistic regression model and another tree-model, which also showed the same patterns (data not shown). Our study has also limitations, however. First, we used data of an elderly ICU population. We believe however that our findings would extrapolate to a general adult ICU population. We chose this subpopulation for the following reasons: 1) the model was already published [10], and 2) because we have access to the data of the entire period (developmental and temporal validation). Second, we did not attempt second level recalibration of the models, which implies adapting each item of the SAPS-II score and changing the structure of the tree-model.

Our findings provide important insights in how SMRs based on parametric and non-parametric models respond to a changing environment (e.g. treatments, new technologies)
and case-mix [20–22]. In our earlier study [8], we found that observed mortality was stable over time, while there was an increasing trend in the patients’ severity of illness. This indicates changes in the patient population and suggests an overall improvement in the provided quality of care of the ICUs as a whole. Although the tree-model did not pick up on these changes, the SMR of the logistic regression model gradually decreased, and eventually unstable (which indicates that the logistic regression model has a shorter “shelf-life” [23]). Importantly, the instability of the global SMR of a model is not necessarily bad; it is an indication that the model is sensing change and requires recalibration. We believe that for comparative audit, where ICUs are compared among each other at a given time-point, it is better to use a more sensitive model that needs to be updated frequently, than using a more stable model that is unable to timely detect change (such as the global improvement in the quality of care over time). Different prognostic models yield different interpretations of the SMR and the interpretation of the delivered quality of care depends on which model is chosen [24].

The differences between the tree-model and the logistic regression model can be explained as follows. First, the tree-model uses different thresholds for the included variables and each of them has only two categories, while the included variables in the logistic regression models are continuous or categorical with multiple categories. This may explain why it takes the tree-model a longer time to adjust its expected mortality in a changing environment (as the probabilities will only alter when the changes in e.g. the physiologic parameters are large enough for patients to fall in the “other” category of a split). Second, the models are based on a different set of variables (for example, Glasgow Coma Scale is used, as such, as the root of the tree-model while it is not explicitly present as a separate covariate in the logistic regression model).

The SPC based strategy proposed in this work provides important insight in SMR behavior over time, which is not otherwise obtainable with traditional statistical methods. Moreover, PreControl Charts provide a systematic way to distinguish between genuine change (e.g. worsening) from inherent noise by using the yellow and red zones and the associated inference rules. Although recalibration is strictly necessary when an out of control situation occurs, we would suggest to use warning signs (e.g. when one point falls either in the yellow zone or is significantly above or below 1) to trigger recalibration.

References